The limits of fair medical imaging AI in real-world generalization

Year: 2024 | Citations: 134 | Authors: Yuzhe Yang, Haoran Zhang, J. Gichoya, Dina Katabi, Marzyeh Ghassemi

Abstract

As artificial intelligence (AI) rapidly approaches human-level performance in medical imaging, it is crucial that it does not exacerbate or propagate healthcare disparities. Previous research established AI's capacity to infer demographic data from chest X-rays, leading to a key concern: do models using demographic shortcuts have unfair predictions across subpopulations? In this study, we conducted a thorough investigation into the extent to which medical AI uses demographic encodings, focusing on potential fairness discrepancies within both in-distribution training sets and external test sets. Our analysis covers three key medical imaging disciplines—radiology, dermatology and ophthalmology—and incorporates data from six global chest X-ray datasets. We confirm that medical imaging AI leverages demographic shortcuts in disease classification. Although correcting shortcuts algorithmically effectively addresses fairness gaps to create 'locally optimal' models within the original data distribution, this optimality is not true in new test settings. Surprisingly, we found that models with less encoding of demographic attributes are often most 'globally optimal', exhibiting better fairness during model evaluation in new test environments. Our work establishes best practices for medical imaging models that maintain their performance and fairness in deployments beyond their initial training contexts, underscoring critical considerations for AI clinical deployments across populations and sites. When tested across tasks, diseases and imaging modalities, the performance of AI models depends on encoding of demographic shortcuts, and correcting for them decreases their ability to generalize in new populations.